Environmental Health Perspectives Supplements
Vol. 102 Suppl 1:95–101 1994

Statistical Issues on the No-Observed-Adverse-Effect Level in Categorical Response

by Takashi Yanagawa[1], Yasuki Kikuchi[2], and Kenneth G. Brown[3]

The determination of the value of the no-observed-adverse-effect level (NOAEL) when observed responses can be categorized by severity (categorical data) and sample sizes are small is discussed. The common situation of only two categories, where only the presence or absence of an effect is observed, is addressed first (dichotomous data). Three tests for dichotomous data are critically examined, including the Brown–La Vange test, a modified version of that test, and Dunnett's multiple comparison test. Although the modified test is an improvement, all three procedures have shortcomings in determining the value of the NOAEL, particularly when the sample size is small. An alternative method is suggested, based on the Akaike information criterion (AIC), which performs well. This method is extended to severity data with an arbitrary number of categories. Use of a dose–response curve for the NOAEL is discussed.

Introduction

As used here, the no-observed-adverse-effect level (NOAEL) is the highest experimental dose at which there is no statistically significant increase in an adverse toxicological end point. This definition restricts the possible values of the NOAEL to the experimental dose values, the only dose levels at which there are observations. Sometimes a dose–response curve is fit to the data, which provides a way of estimating the NOAEL as the lowest dose corresponding to the point on the curve at which the predicted response equals the control rate plus a specified value equal to an acceptable level of increased risk. At low-dose levels, the NOAEL dose may be sensitive to the choice of the dose–response curve fit to the data, particularly in small samples. Consequently, this approach has been suggested for determining the "benchmark dose" as an alternative to the NOAEL, a lower confidence limit to a dose producing some predetermined increase in response rate that will not involve extrapolation far below the experimental range (*1*). The concept of the NOAEL is central to assessment of risk from systematic toxicants, as currently practiced. Inclusion of the NOAEL value in reported laboratory experiments is recommended by the Pharmaceutical Affairs Bureau, Japanese government (GLP, 1989). The U.S. Environmental Protection Agency (EPA) uses the NOAEL in setting regulatory levels for exposure to noncancerous toxic substances (*2,3*).

If d_0 denotes the control dose, and $d_1, d_2, \ldots, d_k$ are increasing dose levels, then the correct choice for the NOAEL is the highest dose value at which the increase in the true risk over the background rate is zero or otherwise acceptably small. One statistical approach that may be used fot the NOAEL is to test the hypothesis of no difference in the true response rates between the control group and a treatment group, pairing the control group for a test with each treatment group sequentially. Williams' test functions this way and can be applied when the data are assumed to be sampled from a normal distribution, e.g., when response is weight gain (*4*). A nonparametric version of that test for use when data are from a continuous but non-normal distribution is described by Shirley (*5*) and Williams (*6*). These tests are order restricted, incorporating *a priori* knowledge that the expected response does not decrease (or increase) as dose level increases. We are unaware of any test in this class for categorical data applicable when severity of response is recorded. For simple dichotomous data (two categories), the test of Brown–La Vange (*7*) and a modified version of that test described here are examples of order-restricted conditional tests. For dichotomous responses, considerable attention has been focused on applying dose–response curves for both cancer and noncancer responses. Crump (*1*) essentially converted his multistage model for cancer data to noncancer application by adding a parameter for a "threshold" dose.

In this paper we are interested in the NOAEL for categorical data, including dichotomous data as a simple case (k=2), from the statistical point of view. Issues related to regulatory applications, such as the use of safety factors with the NOAEL, are

[1]Department of Mathematics, Kyushu University, Fukuoka 812, Japan.
[2]Sasebo College of Technology, Sasebo 857-11, Japan.
[3]Kenneth G. Brown Inc., P.O. Box 16608, Chapel Hill, NC 27516-6608.
Address reprint requests to T. Yanagawa, Department of Mathematics, Kyushu University 33, Fukuoka 812, Japan.
This paper was presented at the International Biostatistics Conference on the Study of Toxicology that was held May 13-15, 1991, in Tokyo, Japan.

not discussed. We study first the behavior of three tests with dichotomous data, including the Brown–La Vange method, a modification of this method, and the multiple comparison test of Dunnett. It is shown that, although the modified test is an improvement over the other two tests, all three tests have serious shortcomings when the sample size is small. A new test implementing the Akaike information criterion (AIC) (*8*) is shown to work well. The AIC test is generalized to an arbitrary number of categories for application with severity data. Finally, application of the AIC with a dose–response model for noncancer end points is outlined, to be more fully developed in a follow-up paper (Yanagawa et al., unpublished data).

Tests for Dichotomous Response Data

An experiment with dichotomous response data is described by the number of experimental subjects at risk (n_i), the number with the response of interest (r_i), and the exposure level (d_i), for $i = 0,1,\ldots, k$. The subscript zero refers to the control group, making $d_0 = 0$; otherwise the dose values are arbitrary, subject to order $0 = d_0 < d_1 < \ldots < d_k$. The true, but unknown response rate at dose d_i is denoted by p_i, $i = 0,1,\ldots, k$. It is assumed that the samples are random and mutually independent, and that the number of responses r_i at d_i is binomially distributed with parameters (n_i,p_i), $i = 0,1,\ldots,k$. It is also assumed to be known *a priori* that the true response rate is nondecreasing as dose increases, i.e., $0 \leq p_0 \leq p_1 \leq \ldots \leq p_k \leq 1$. Alternatively, one could assume that $1 \geq p_0 \geq p_1 \geq \ldots \geq p_k \geq 0$.

Let δ^* denote the largest d_i value such that $p_0 = p_i$. The test procedure to be described is a method by which to assign the NOAEL a dose value based on the sample data, conditional on the total number of responses observed over all dose groups, namely, $S(r) = (r_0 + r_1 + \ldots + r_k)$. In the following section, we describe the Brown–La Vange (BLV) test, a modified form of it (MBLV), and the Dunnett-type multiple comparison test (DMC). The tests are compared when $k = 2$ for simplicity.

Brown–La Vange Test

Without the constraint $p_0 \leq p_1 \leq \ldots \leq p_k$, the maximum likelihood estimate (MLE) of p_i is r_i/n_i. The MLE of p_i under the order restriction, however, is m_i/n_i, where m_i is constructed by the pool-adjacent-violators algorithm (*9,10*). The BLV step-up tests are based on the values of $(m_0, m_1, \ldots, m_k)$, as described for $k = 2$ in the following. Initially, the null hypothesis H_o^1: $p_0 = p_1$ is tested against H_a^1: $p_0 < p_1$. If H_o^1 is rejected, then the NOAEL takes the value d_0; if it is not rejected, then the NOAEL is d_1 or d_2, as determined by the subsequent test. Thus we could write H_o^1: $\delta^* = d_1$ or d_2 versus H_a^1: $\delta^* = d_0$. Let $t_1 = m_1/n_1 - m_0/n_0$ be the test statistic for H_o^1. For a specified test size, α_1, reject H_o^1 if t_1 takes a value as large as k_1, where k_1 is the smallest constant such that $Pr[t_1 \geq k_1 | S(r)] \leq \alpha_1$, when H_o^1 is true. Here $|S(r)$ should be read as "conditional on $S(r) = (r_0 + r_1 + r_2)$." If H_o^1 is rejected, then the NOAEL takes the value d_0. If H_o^1 is not rejected, then H_o^2: $p_0 = p_2 \,|\, p_0 = p_1$ should be tested, where $|p_0 = p_1$ should be read as "conditional on having not rejected H_o^1: $p_0 = p_1$." The alternative hypothesis can be written as H_a^2: $p_o < p_2 | p_0 = p_1$. Equivalently, the second test is of H_o^2: $\delta^* = d_2 | \delta^* > d_0$ versus H_a^2: $\delta^* = d_1 | \delta^* > d_0$. If H_o^2 is rejected, then the NOAEL takes the value d_1; otherwise, it takes the value d_2. For a specified test size, α_2, the test rejects H_o^2 if $t_2 = m_2/n_2 - m_0/n_0 \geq k_2$, where k_2 is the smallest constant such that

$$Pr(t_2 \geq k_2 \mid S(r),\, t_1 < k_1) = \frac{Pr(t_1 < k_1,\, t_2 \geq k_2 \mid S(r))}{Pr(t_1 < k_1 \mid S(r))} \leq \alpha_2 \qquad (1)$$

under H_o^2.

Dunnett-Type Multiple Comparison Test

Alternatively, we may apply the Dunnett multiple comparison test (DMC) for the NOAEL based on the adjusted response. For a specified test size, α, this test first selects the smallest constant k such that

$$Pr(t_2 < k \mid S(r)) \geq 1 - \alpha \qquad (2)$$

under $p_0 = p_1 = p_2$, from which the NOAEL is determined according to: if $t_1 \geq k$, then NOAEL $= d_0$; if $t_1 < k$ and $t_2 \geq k$, then NOAEL $= d_1$; if $t_1 < k$ and $t_2 < k$, then NOAEL $= d_2$.

Modification to the Brown–La Vange Test

This test pools the responses at d_0 and d_1, if no significance difference is detected between these dose levels to increase the power of the test. The test is based on the values of $(r_0, r_1, \ldots, r_k)$, the naive responses. The test procedure is the same as that of the Brown–La Vange test except the test statistic. The test statistic for H_o^1 is $u_1 = r_1/n_1 - r_0/n_0$, and the test statistic for H_o^2 is $u_2 = r_2/n_2 - (r + r_1)/(n_0 + n_1)$. For a specified test size, α_1, the test rejects H_o^1 if $u_1 \geq k_1^*$, where k_1 is the smallest constant such that $Pr[u_1 \geq k_1^* | S(r)] \leq \alpha_1$, when H_o^1 is true. For a specified test size, α_2, the test rejects H_o^2 if $u_2 \geq k_2^*$, where k_2^* is the smallest constant such that

$$Pr(u_2 \geq k_2^* \mid S(r),\, u_1 < k_1^*) \leq \alpha_2 \qquad (3)$$

under H_o^2.

Small-Sample Behavior of the Tests

We compare the tests in detail when $k = 2$, $n_0 = n_1 = n_2 = 10$, and $S(r) = 4$. When $S(r) = 4$ is given, the number of all possible configurations of the tables of $n_0 = n_1 = n_2 = 10$ is 15, as shown in Table 1. The probability of each entry in the table, when $p_0 = p_1 = p_2$, has been computed from a multiple hypergeometric distribution and included in the table. Consequently, the probability is the chance occurrence of an entry in the absence of an effect.

The distributions of statistics t_1 and u_1 have been tabulated from the entries in Table 1 and are displayed in Table 2. Table 2 shows that, in the case of the conditional test based on the adjusted response, the values of the test statistic t_1 take only four points with positive probability, and the jumps of the cumulative

Table 1. List of all feasible tables when $n_0 = n_1 = n_2 = 10$ and S)r) = 4.

	Entry Number														
Response	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
r_0	0	0	0	0	0	1	1	1	1	2	2	2	3	3	4
r_1	0	1	2	3	4	0	1	2	3	0	1	2	0	1	0
r_2	4	3	2	1	0	3	2	1	0	2	1	0	1	0	0
Probability	0.0077	0.0438	0.0739	0.0438	0.0077	0.0438	0.1642	0.1642	0.0438	0.0739	0.1642	0.0739	0.0438	0.0438	0.0077

Table 2. The conditional distributions of the statistics t_1 and u_1 conditioned on $S(r) = 4$.

	Value of t_1			
	0.00	0.05	0.10	0.20
Cumulative prob.	1	0.3771	0.1691	0.1253

	Value of u_1								
	−0.40	−0.30	−0.20	−0.10	0.00	0.10	0.20	0.30	0.40
Cumulative prob.	1	0.9923	0.9486	0.8309	0.6229	0.3771	0.1691	0.0515	0.0077

probability are so large that no finite k_1 exists when the values of α_1 are specified less than 0.1253. Similarly, the Dunnett-type multiple comparison test does not select d_0 as the NOAEL when the values of α are specified to be less than 0.221. The modified test is also a conditional test, but based on the naive response, and the statistic u_1 takes more values than t_1, and the jumps of the cumulative probabilities are relatively small. Thus we may test H_o^1 at test sizes less than 10%, e.g., at $\alpha_1 = 0.0515$ or 0.0077.

Table 3. The entries in Table 1 that select d_0, d_1, and d_2 as the NOAEL when the BLV test, Dunnet type test the MBLV test, and the AIC are applied.

	Test size				NOAEL		
Test	α_1	(k_1)	α_2	(k_2)	d_0	d_1	d_2
MBLV	0.05[a]	(.40)	0.05[a]	(0.40)	5	1	AO
			0.10[a]	(0.25)	5	1, 2, 6	AO
	0.10[a]	(0.30)	0.05[a]	(0.40)	4, 5	1	AO
			0.10[a]	(0.40)	4, 5	1	AO
	0.0515	(0.30)	0.0081	(0.40)	4, 5	1	AO
			0.1004	(0.25)	4, 5	1, 2, 6	AO
	0.1691	(0.20)	0.0092	(0.40)	3, 9, 4, 5	1	AO
			0.1146	(0.25)	3, 9, 4, 5	1, 2, 6	AO
BLV	0.05[a]	(—[b])	0.05[a]	(0.40)	None	1	AO
			0.10[a]	(0.25)	None	1, 2, 6	AO
	0.10[a]	(—[b])	0.05[a]	(0.40)	None	1	AO
			0.10[a]	(0.25)	None	1, 2, 6	AO
	0.1253	(0.20)	0.0088	(0.40)	3, 4, 5	1	AO
			0.0588	(0.30)	3, 4, 5	1, 2	AO
			0.1089	(0.25)	3, 4, 5	1, 2, 6	AO
Dunnett-type	$\alpha = 0.05$[a] $(k = 0.40)$				None	1	AO
	$= 0.10$[a] $(k = 0.25)$				None	1, 2, 6	AO
	$= 0.20$[a] $(k = 0.25)$				None	1, 2, 6	AO
	$= 0.221$ $(k = 0.20)$				3, 4, 5	1, 2, 6	AO
AIC					3, 4, 6	1, 2, 6	AO

Abbreviations: NOAEL, no-observed-adverse-effect level; BLV, Brown–La Vange test; MBLV, modified Brown–La Vange test; AIC, Akaike information criterion; AO, all others.

The three tests are applied to each entry in Table 1 with the results summarized in Table 3. When entry no. 4 or 5 is observed (Table 1), the modified test selects d_0 as the NOAEL at the test size $\alpha_1 = 0.10$ and $\alpha_2 = 0.1004$; when entry no. 1, 2, or 6 is observed (Table 1), then d_1 is selected as the NOAEL; and when any other number is observed, d_2 is selected as the NOAEL. The probabilities of the correct decision for MBLV under $p_0 = p_1 \leq p_2$ (case 2) and $p_0 \leq p_1 = p_2$ (case 3) may be computed using the formula

$$Pr(u_1 < k_1^*, u_2 \geq k_2^* \mid S(r)) = Pr(u_1 < k_1^* \mid S(r)) Pr(u_2 \geq k_2^* \mid S(r), u_1 < k_1^*) \qquad (4)$$

by specifying the values of p_0 and the values of the added risk $p_2 - p_0$. Figure 1 shows the probabilities of the correct decision at the test size $\alpha_1 = 0.0515$ and $\alpha_2 = 0.1004$ for the values of $p_0 = 0.05$ and 0.15, and $p_2 - p_0 = 0.05$, 0.30 (0.05) in cases 2 and 3. The figure shows that the probabilities are relatively large when $p_0 = 0.05$ in case 2, but small when $p_0 = 0.15$ in case 3. For example, when $p_0 = 0.15$, the probability of the correct decision is only 0.182 in case 3, even if the added risk is 0.30. Consequently, the power of the test to detect an effect depends on the background rate p_0, as well as on the added risk.

Summary: Flaws of the Statistical test

The findings from Tables 2 and 3 and Figure 1 are summarized as follows: *a*) The BLV test failed to select d_0 as the NOAEL at the routine test size, i.e., $\alpha_1 = 0.05$ or 0.10. The same is observed for the DMC test at $\alpha = 0.10$ or 0.20. *b*) For a step-up test, such as the BLV, the influence of the first step is considerable. The key is in the selection of the value of α_1. For example, the probabilities of the correct decision by the BLV (and the DMC as well) is zero in case 3 at the test sizes $\alpha_1 = 0.10$ ($\alpha = 0.20$), even when the added risk is 0.30, because of the reason stated above. It is apparent from Table 3 that if we specify $\alpha_1 = 0.1253$, the behavior of the BLV test is much improved. The problem is that it is not easy to determine the test size to use. *c*) The DMC test is not a step-up test, but has a similar property to the BLV test. Generally, if sample sizes are small and a test is constructed based on the adjusted responses, then the jumps in the values of the tail probabilities are remarkable, frequently larger than 0.05. It is not justifiable in those situations to carry out a test with a routine test size of 0.05. *d*) The modified test (MBLV) removes the difficulty due to the first step and performs better than the BLV or DMC test. With small sample sizes, however, the probability of the correct decision in case 3 is disappointingly small. *e*) A puzzling aspect of the modified test may be noted. Suppose that entry no. 4 in Table 4 is observed. If we set $\alpha_1 = 0.05$ and $\alpha_2 = 0.10$, then Table 3 shows that d_2 is selected as the NOAEL, but if

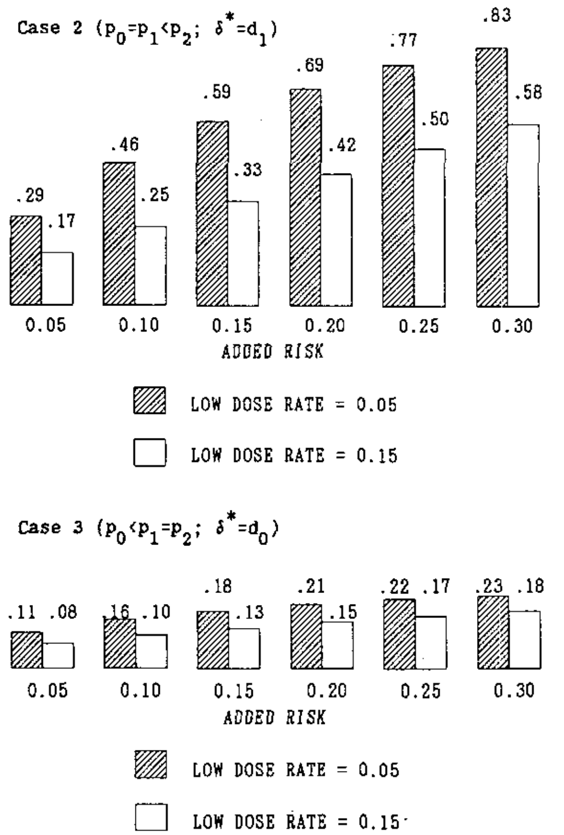

FIGURE 1. Probability of correct decision by the MBLV test.

$\alpha_1 = 0.0515$ and $\alpha_2 = 0.1004$, then d_0 is selected as the NOAEL. The selection order of the values d as the NOAEL would reasonably follow the pattern $d_2 \rightarrow d_1 \rightarrow d_0$, instead of jumping from d_2 to d_0. The same phenomenon occurs with BLV and DMC. *f*) We have applied the three tests to other small-sample tables and have observed that the smaller the sample sizes, the larger the values selected as the NOAEL. This behavior, discussed by Crump (*1*) and others, is unacceptable because smaller samples tend to make the dose levels appear safer. Brown and Erdreich (*7*) emphasized calculation of statistical power to detect an effect level of interest before drawing a conclusion. Those calculations, however, are cumbersome. A preferred approach may be to consider jointly the test size and sample size. It is not easy to develop this idea in the framework of statistical testing, but it can be achieved in the framework of model selection. We explore the use of the Akaike information criterion (AIC) for this objective in the next section.

Application of the AIC

We continue with the same notation and conditions described in the previous sections, i.e., $k = 2$ with dichotomous data. Let

$$\gamma_1 = \log\left(\frac{p_1(1-p_0)}{(1-p_1)p_0}\right), \quad \gamma_2 = \log\left(\frac{p_2(1-p_0)}{(1-p_2)p_0}\right). \quad (5)$$

The parameters γ_1 and γ_2 are the log odds ratios of the effect at d_1 and d_2, respectively, relative to the effect at d_0. Note that

$p_0 = p_1 = p_2$ if and only if $\gamma_1 = \gamma_2 = 0$
$p_0 = p_1 \le p_2$ if and only if $\gamma_1 = 0, \gamma_2 \ge 0$
$p_0 \le p_1, p_0 \le p_2$ if and only if $\gamma_1 \ge 0, \gamma_2 \ge 0$

and that the order restriction $p_0 \le p_1 \le p_2$ is equivalent to $\gamma_1 \ge 0$, and $\gamma_2 - \gamma_1 \ge 0$.

The conditional log likelihood conditioned on $S(r) = (r_0+r_1+r_2)$ is

$$l(\gamma_1, \gamma_2) = const + \gamma_1 r_1 + \gamma_2 r_2$$

$$- \log \sum{}^{*} \binom{S(r)}{x_1, x_2}\binom{n - S(r)}{n_1 - x_1, n_1 - x_2} \exp(\gamma_1 x_1 + \gamma_2 x_2), \quad (6)$$

were

$$\binom{s}{x_1, x_2} = \frac{s!}{(s - x_1 - x_2)!\, x_1!\, x_2!}, \quad (7)$$

and Σ^* is the summation that extends over all integers x_1 and x_2 such that $n_1 \ge x_1 \ge 0$, $n_2 \ge x_2 \ge 0$ and $S(r) - x_1 - x_2 \ge 0$.

Put $L(\gamma_1, \gamma_2) = 2\,l(\gamma_1, \gamma_2) - 2$ (number of parameters involved in the likelihood), which is the likelihood function penalized by the number of parameters involved in the model. Let $\hat\gamma_1$ and $\hat\gamma_2$ be the maximum likelihood estimators (MLEs) of γ_1 and γ_2 which maximize $L(\gamma_1, \gamma_2)$. Then $L(\hat\gamma_1, \hat\gamma_2)$ measures the goodness of fit of the two-parameter model to the data. In the present setup, the exact fit to the data is achieved by the two-parameter model because the number of degrees of freedom is two. Next we suppose that $\gamma_1 = 0$ is known and that γ_2 is the only parameter in the model. Let $\hat\gamma_2^*$ be the MLE of γ_2. Then $L(\gamma_1 = 0, \hat\gamma_2^*)$ is a measure of the goodness of fit of the one-parameter model to the data. Of course, this model does not provide an exact fit because it involves only one parameter. The penalized likelihood has been established to measure the goodness of fit by adjusting the number of the parameters involved in the model. Thus if $L(\hat\gamma_1, \hat\gamma_2) < L(\hat\gamma_1 = 0, \hat\gamma_2^*)$, we may select the one-parameter model. This idea of the model selection is first proposed by Akaike (*8*) and is widely known as the Akaike information criterion.

We take into account the order restriction $\gamma_1 \ge 0$, and $\gamma_2 - \gamma_1 \ge 0$ and apply the AIC for the determination of the NOAEL as follows:

a) If $\hat\gamma_1 > 0$ and $\hat\gamma_2 - \hat\gamma_1 > 0$, then compare $L(\hat\gamma_1, \hat\gamma_2)$, $L(\gamma_1 = 0, \hat\gamma_2^*)$, and $L(\gamma_1 = 0, \gamma_2 = 0)$.
If $L(\hat\gamma_1, \hat\gamma_2)$ is the largest, NOAEL $= d_0$,
if $L(\gamma_1 = 0, \hat\gamma_2^*)$ is the largest, NOAEL $= d_1$,
if $L(\gamma_1 = 0, \gamma_2 = 0)$ is the largest, NOAEL $= d_2$.

b) If $\hat\gamma_1 > 0$ and $\hat\gamma_2 - \hat\gamma_1 \le 0$, then put $\gamma = \gamma_1 = \gamma_2 L$ and obtain the MLE $\hat\gamma$ of γ.
If $\hat\gamma \le 0$, NOAEL $= d_2$.
If $\hat\gamma > 0$, then compare $L(\gamma_1 = 0, \gamma_2 = 0)$ with $L(\gamma_1 = \hat\gamma, \gamma_2 = \hat\gamma)$:
if $L(\gamma_1 = \hat\gamma, \gamma_2 = \hat\gamma)$ is the largest, decide NOAEL $= d_0$,
if $L(\gamma_1 = 0, \gamma_2 = 0)$ is the largest, decide NOAEL $= d_2$.

c) If $\hat\gamma_1 \le 0$ and $\hat\gamma_2 - \hat\gamma_1 > 0$, then obtain $\hat\gamma_2^*$, the MLE of γ_2.
If $\hat\gamma_2^* \le 0$, decide NOAEL $= d_2$.
If $\hat\gamma_2^* > 0$, then compare $L(\gamma_1 = 0, \gamma_2 = 0)$ with $L(\gamma_1 = 0, \gamma_2 = \hat\gamma_2^*)$:
if $L(\gamma_1 = 0, \gamma_2 = \hat\gamma_2^*)$ is the largest, decide NOAEL $= d_1$,
if $L(\gamma_1 = 0, \gamma_2 = 0)$ is the largest, decide NOAEL $= d_2$.

d) If $\hat\gamma_1 \le 0$ and $\hat\gamma_1 - \hat\gamma_1 \le 0$, then decide NOAEL $= d_2$.

This procedure is applied to the entries in Table 1. The results are given in the last row of Table 3. Figure 2 illustrates the probability of the correct decision for case 2 and for case 3. Comparing these results with the outcomes of the preceding tests, one can clearly see the superiority of this method. In particular, the AIC method relieves the problem of selecting the test size described earlier and increases the probability of a correct decision.

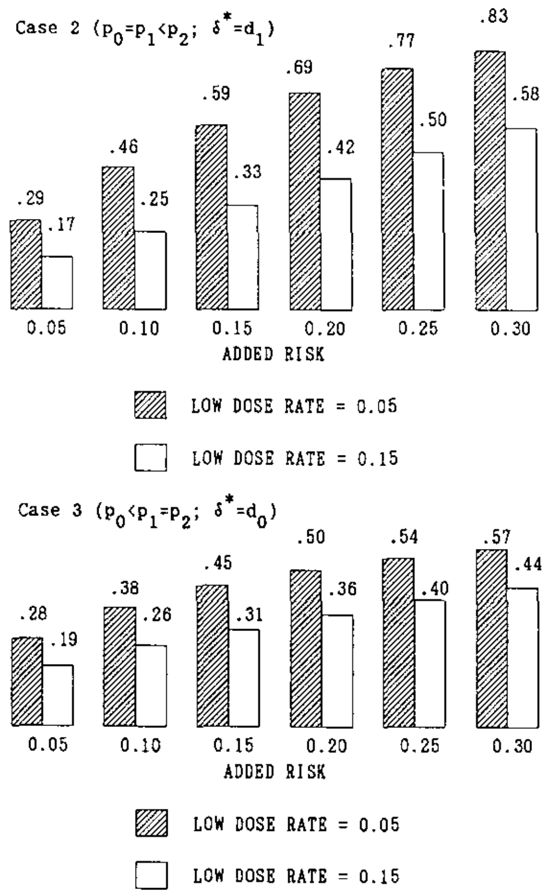

FIGURE 2. Probability of correct decision by the AIC.

Extension of the AIC for a NOAEL in Categorical Data

We extend application of the AIC to determination of the NOAEL in categorical response data. Suppose that there are $b+1$ categories, and let r_{ij} be the number of responses in the jth category at exposure level d_i, $i = 0,1,\ldots,k$; $j = 0,1,\ldots,b$. Let p_{ij} be the response probability at the ith exposure level and jth category. It is assumed that the samples are random and mutually independent and that the response at dose $i(r_{i0}, r_{i1},\ldots,r_{ib})$ are multinomially distributed with parameters $(n_i, p_{i0}, p_{i1},\ldots,p_{ib})$, $i = 0,1,\ldots,k$. Let $C_0 \leq C_1 \leq \ldots \leq C_b$ be given scores that are assigned to the categories. For example, we might assign $C_0 = 0$, $C_1 = 1,\ldots,C_b = b$, or alternatively, the Wilcoxon score could be assigned. We introduce the following model for the response probabilities:

$$\log\left(\frac{p_{ij}}{p_{i0}}\right) = \beta_i (C_j - C_0), \quad j=1,2,\ldots,b;\ i=0,1,\ldots,k \tag{8}$$

It is assumed to be known *a priori* that $\beta_0 \leq \beta_1 \leq \ldots \leq \beta_k$. Alternatively, one could assume that $\beta_0 \geq \beta_1 \geq \ldots \geq \beta_k$. This assumption generalizes the previous assumption regarding the order restriction of the response probabilities.

Put $S(r_j) = r_{0j} + r_{1j} + \ldots + r_{kj}$. The conditional log likelihood of $\{r_{ij}\}$ conditioned on $S(r_j)$, $j = 0, 1,\ldots,b$, is given by:

$$l(\gamma_1, \gamma_2,\ldots,\gamma_k) = const + \sum_{i=1}^{k} \gamma_i \sum_{j=1}^{b} r_{ij}(C_j - C_0)$$

$$- \log \sum{}^{*} \prod_{j=0}^{b} \frac{S(r_{.j})!}{x_{0j}!x_{1j}!\ldots x_{kj}!} \exp\left[\sum_{i=1}^{k} \gamma_i \sum_{j=1}^{b} x_{ij}(C_j - C_0)\right], \tag{9}$$

where $\gamma_i = \beta_i - \beta_0$ and Σ^* is the summation that extends over all combinations of the integers $\{x_{0j}, x_{1j},\ldots,x_{kj}\}$ such that $n_i \geq x_{ij} \geq 0$ and $x_{0j}+x_{1j}+\ldots+x_{kj} = S(r_j)$, $j = 0,1,\ldots,b$. The log likelihood shows that it is sufficient to carry out the statistical inference on $\gamma_1, \gamma_2\ldots$, and γ_k based on statistics

$$T_i = \sum_{j=1}^{b} r_{ij}(C_j - C_0), \quad i = 1, 2,\ldots, k \tag{10}$$

The model (8), which seems somewhat artificial, is a mathematical device to lead to this reasonable result.

The order restriction is represented by $\gamma_1 \geq 0$, and $\gamma_i - \gamma_{i-1} \geq 0$, $i = 2,3,\ldots,k$. The AIC is applied for the determination of the NOAEL taking this restriction into account. for $k = 2$, the procedure is the same as that given in the preceding section.

Use of a Dose-Response Curve

We define

$$E\left(\frac{T_i}{n_i}\right) = \sum_{j=1}^{b} (C_j - C_0)\, p_{ij} \tag{11}$$

as an average dose response. Fitting a smooth curve

$$h(\theta{:}d) = \theta_1 d + \theta_2 d^2 + \ldots + \theta_p d^p \ (p<k)$$

to $\gamma_1, \gamma_2,\ldots,\gamma_k$, the average dose response curve is represented by

$$p(d{:}h_\theta, \beta_0) = \sum_{j=1}^{b} \{(C_j - C_0) q_j(\theta{:}d) / [1 + \sum_{l=1}^{b} q_l(\theta{:}d)]\}, \tag{12}$$

where

$$q_j(\theta{:}d) = \exp\{(C_j - C_0)[\beta_0 + h(\theta{:}d)]\}. \tag{13}$$

Similar to the preceding section, the AIC may be applied to the conditional distribution to select the optimum value of p that fits the data best and to obtain the conditional MLE $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p$, of $\theta_1, \theta_2, \ldots, \theta_p$. It is not feasible to estimate β_0 from the conditional likelihood function. One way to estimate it is to use the full likelihood function, assuming that $h(\hat{\theta}:d)$ is known. The other method is to use the data in the control group whose response probability contains only β_0. It is not easy to get the variance of $\hat{\beta}_0$, but the approximate variance of $h(\hat{\theta}:d)$ is readily available from the Fisher information of the conditional distribution. Thus, in this paper, we ignore the variation of β_0 for illustrative purposes and only take into account the uncertainty of estimating $\hat{\theta}$. Let UB (d) be the 95% upper confidence bound of $h(\hat{\theta}:d)$. Then for a given constant, c, the NOAEL = d^* may be found by solving either

$$c = \frac{p(d^*:UB(d^*), \hat{\beta}_0) - p(0:0, \hat{\beta}_0)}{1 - p(0:0, \hat{\beta}_0)} \quad \text{(relative risk)} \quad (14)$$

or

$$p(0:0, \hat{\beta}_0) + c = p(d^*: UB(d^*), \hat{\beta}_0) \quad \text{(additive risk)}. \quad (15)$$

As in Crump (*1*), we may introduce a threshold factor. That extension, and the construction of a reliable confidence interval, will be discussed in a follow-up paper.

An Application

Fitzhugh et al. (*11*) report results of exposing Osborne-Mendel rats for 2 years to diets containing aldrin in 0, 0.5, 2, 10, 50, 100, and 150 ppm. The study reports the degree of liver changes categorized as none, trace, very slight, slight, slight/moderate to moderate, and greater than moderate. For the purpose of illustration, we use a part of data as shown in Table 4. The scores are assigned as $C_0=0$, $C_1=1$, $C_2=2$, $C_3=3$, and then the AIC procedure is applied. The conditional MLE of γ_1, γ_2 and γ_3 are obtained as $\hat{\gamma}_1 = 1.193$, $\hat{\gamma}_2 = 2.107$, $\hat{\gamma}_3 = 2.538$. These estimates satisfy the order restrictions $\hat{\gamma}_1 \geq 0$, $\hat{\gamma}_2 - \hat{\gamma}_1 \geq 0$, and $\hat{\gamma}_3 - \hat{\gamma}_2 \geq 0$. The values of the penalized likelihood are given by

$L(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3) = -26.14$, $L(\gamma_1 = 0, \gamma_2 = \hat{\gamma}_2^*, \gamma_3 = \hat{\gamma}_3^*) = -25.63$

$L(\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = \hat{\gamma}_3^{\#}) = -30.42$, $L(\gamma_1 = 0, \gamma_2 = 0, \gamma_3 = 0) = -37.87$,

where $\hat{\gamma}_2^* = 1.287$, $\hat{\gamma}_3^* = 1.715$ are the MLEs under $\gamma_1 = 0$, and $\hat{\gamma}_3^{\#} = 0.991$ is the MLE of γ_3 under $\gamma_1 = 0$, $\gamma_2 = 0$. The second likelihood provides the maximum among the four choices, so d_1 is chosen as the NOAEL.

We also extended the modified test (MBLV) to apply to these data for comparison. The test leads to d_2 as the NOAEL for $\alpha_1 = 0.10$ and $\alpha_2 = 0.10$. The dose-response curve method is also applied for comparison, particularly because it is not restricted to experimental dose values for choice of the NOAEL.

Table 4. Degree of liver changes.

Dose	Severity[a] N	T	VS	S	Total
0	16	1	0	0	17
0.5	15	4	0	0	19
2	10	8	0	1	19
10	11	3	7	1	22

[a]N, none; T, trace; VS, very slight; S, more than slight.

The AIC selects $h(\theta:d) = \theta_1 d + \theta_2 d^2$ with $(\hat{\theta}_1, \hat{\theta}_2) = (0.9778, -0.0772)$. The variance and covariance of $(\hat{\theta}_1, \hat{\theta}_2)$ are $V(\hat{\theta}_1) = 0.1613$, $V(\hat{\theta}_2) = 0.00123$ and $\mathrm{cov}(\hat{\theta}_1, \hat{\theta}_2) = -0.0140$. Assuming $(\hat{\theta}_1, \hat{\theta}_2)$ is known, the estimate of β_0 from the full likelihood is -3.0132. Alternatively, the estimate of β_0 from the control group data alone is -2.8898. Figure 3 shows the average dose-response curve and its upper 95% confidence bound. we may assess the NOAEL from Figure 3. For example, when $c = 0.2$ is specified in the relative risk model, the NOAEL is assessed to be 0.85.

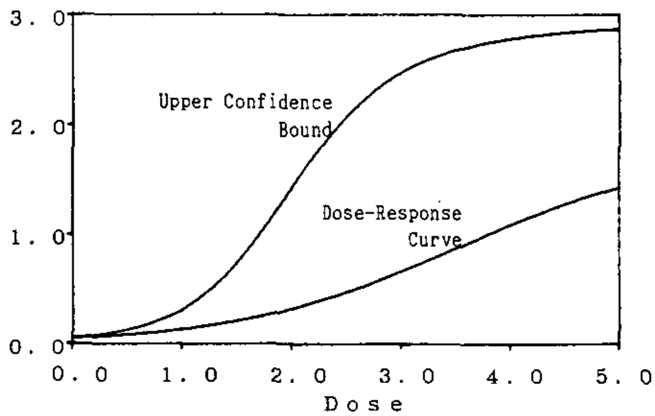

FIGURE 3. Dose-response curve.

Discussion

We have developed several methods of selecting the NOAEL when the responses are measured by severity and also when the sample sizes are small. Our conclusions are as follows: *a*) If one wants to select the NOAEL from the experimental dose levels $\{d_0, d_1, \ldots, d_k\}$, then implementation of the AIC in the order restricted likelihood method is preferable to a testing approach, as demonstrated for three alternative test procedures. *b*) If one wants to select the NOAEL from the full experimental range of doses, (d_0 to d_k), then a dose-response curve is required to estimate responses between observed values. The choice of the dose-response curve may affect the outcome, but fitting the "average dose-response curve" as described is reasonable. The choice of c in that model should be chosen carefully. The NOAEL can be based on either relative risk or additive risk, depending on one's objective.

REFERENCES

1. Crump, K.S. A new method for determining allowable daily intakes. Fundam. Appl. Toxicol. 4: 854–871 (1984).
2. U.S. EPA. Interim Methods for Development of Inhalation Reference Doses. EPA/600/8-88/066F, Office of Health and Environmental Assessment. U.S. Environmental Protection Agency, Washington, DC, 1989.
3. U.S. EPA. Reference Dose (RfD): Description and Use in Health Risk Assessment, Appendix A, Integrated Risk Information System (IRIS). Office of Health and Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC, 1986.
4. Williams, D. A. A test for differences between treatment means when several dose levels are compared with a zero dose control. Biometrics 27: 10–117 (1971).

5. Shirley, E. A nonparametric equivalence of Williams' test for contrasting increasing dose levels of a treatment. Biometrics 33: 386–389 (1977).
6. Williams, D. A. A note on Shirley's nonparametric test for comparing several dose levels with a zero-dose control. Biometrics 42: 183–186 (1986).
7. Brown, K. G., and Erdreich, L. S. Statistical uncertainty in the no-observed-adverse-effect level. Fundam. Appl. Toxicol. 13: 235–244 (1989).
8. Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory (B. N. Petrov and F. Csaki, Eds.), Akademiai Kiado, Budapest, 1973, pp. 267–281.
9. Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., and Brunk, H. D. Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression. John Wiley and Sons, New York, 1972.
10. Schoenfeld, D. A. Confidence bounds for normal means under order restrictions, with application to dose response curves, toxicology experiments and low dose extrapolation. J. Am. Stat. Assoc. 81: 186–195 (1983).
11. Fitzhugh, O. G., Nelson, A. A., and Quaife, M. L. Chronic oral toxicity of aldrin and dieldrin in rats and dogs. Food Cosmet. Toxicol. 2: 551–562 (1964).